Regular paper

Assessing the 5S ribosomal RNA heterogeneity in *Arabidopsis thaliana* using short RNA next generation sequencing data

Maciej Szymanski✉ and Wojciech M. Karlowski

Department of Computational Biology, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University in Poznan, Poznań, Poland

In eukaryotes, ribosomal 5S rRNAs are products of multigene families organized within clusters of tandemly repeated units. Accumulation of genomic data obtained from a variety of organisms demonstrated that the potential 5S rRNA coding sequences show a large number of variants, often incompatible with folding into a correct secondary structure. Here, we present results of an analysis of a large set of short RNA sequences generated by the next generation sequencing techniques, to address the problem of heterogeneity of the 5S rRNA transcripts in Arabidopsis and identification of potentially functional rRNA-derived fragments.

INTRODUCTION

5S ribosomal RNA (5S rRNA) is a conserved, ubiquitous component of the large ribosomal subunit. Except for mitochondria of animals and some fungi, it is found in the ribosomes of all Bacteria, Archaea and Eukaryota and is essential for their function (Szymanski *et al.*, 2003).

In Bacteria, Archaea and lower eukaryotes, the expression of 5S rRNAs is directly linked to the expression of other ribosomal RNAs that are processed from a long pre-rRNA transcript. In higher eukaryotes, the 5S rRNA genes are separated from the genes encoding long LSU and SSU rRNAs and 5.8S rRNAs, and are transcribed from clusters of tandemly repeated units consisting of a conserved 5S RNA-coding sequence and a non-transcribed spacer (NTS) region. The number of clusters, repeat units and their distribution on the chromosomes varies between different species. Unlike expression of other rRNA genes transcribed by the RNA polymerase I, transcription of eukaryotic 5S rRNAs depends on the activity of RNA polymerase III and the transcription factor IIIA (TF IIIA) that binds the promoter region localized within 5S rRNA coding sequence (Ciganda & Williams, 2011).

Although RNA sequencing of mature 5S rRNA usually results in a single nucleotide sequence, heterogeneity of 5S rRNA was observed in many organisms from various taxonomic groups (Szymanski *et al.*, 2016). The best studied example of the functionally heterogenous 5S rRNAs is the expression of somatic- and oocyte-type 5S rRNAs differing at five nucleotide positions in *Xenopus* laevis (Ford & Southern, 1973). Somatic-type 5S rRNA is expressed in both, oocytes and somatic cells. On the other hand, expression of the oocyte-type 5S rRNA is restricted to oogenesis and early developmental stages (Peterson *et al.*, 1980). Expression and methylation-dependent regulation of embryonic and somatic types of 5S rRNA was also recently demonstrated in the sea urchin *Paracentrotus lividus* (Dimarco *et al.* 2012; Bellavia *et al.* 2013).

In plants, the first example of 5S rRNA sequence heterogeneity was reported in rice embryos in which, during germination, expression of two 5S rRNA species was observed (Hariharan *et al.*, 1987). A number of different 5S rRNA variant genes within individual species were revealed by sequencing of the plant 5S rRNA repeat units used as markers for phylogenetic analyses (Baum & Johnson, 1997; Baum *et al.* 2013). Although in most cases RNA sequencing suggests existence of a single mature 5S rRNA species, there was a significant sequence variation at the DNA level both, within the 5S rRNA coding sequences, as well as within the NTS regions (Szymanski *et al.*, 1995).

The heterogeneity of 5S rRNA-coding DNA sequences and their transcripts was also systematically analyzed in *Arabidopsis thaliana* by sequencing of the 5S rDNA repeat units and the reverse transcription products of 5S rRNA (Cloix *et al.* 2002).

Here, we present a novel attempt to sample sequence heterogeneity of the 5S rRNA in *Arabidopsis thaliana* by analysis of a large set of short RNA fragments obtained from high throughput experiments using the next generation sequencing techniques. Analysis of the coverage of the 5S rRNA coding sequences by 18–30 nucleotide long sequencing reads suggests that the number of transcribed 5S rRNA species may be greater than has been determined previously. Moreover, we identified a region within the 5S rRNA sequence that is a source of very stable short RNAs.

RESULTS AND DISCUSSION

Variants of 5S rDNA repeats

As in other higher eukaryotes, *Arabidopsis thaliana* 5S rRNA genes are arranged in clusters of tandem repeats consisting of 120 bp-long transcribed sequences and approximately 380 bp-long non-transcribed spacers. The number of repeat units per haploid genome of *A. thaliana* was estimated to be approximately 1000 (Campell

✉e-mail: mszyman@amu.edu.pl
Abbreviations: NTS, non-transcribed spacer; TF IIIA, transcription factor IIIA

et al., 1992). The 5S rDNA repeat clusters are localized within the pericentromeric regions on chromosomes 3, 4 and 5 (Fransz *et al.*, 1998; Murata *et al.*, 1997).

The genomic sequences of all 5S rRNA coding sequences were obtained from the RNA families database Rfam (Nawrocki *et al.*, 2015). A set of 562 unique variants for further analysis was generated from 1347 sequences by filtering out records that were too short (less than 116 bp) or interrupted by RNA polymerase III termination signals (stretches of four or more Ts). The unique sequences of 5S rRNA were aligned to the reference *A. thaliana* 5S rRNA sequence (Barciszewska *et al.*, 1994) using clustalW (Larkin *et al.*, 2007). The most abundant among the genomic sequences identified by Rfam is the sequence identified in earlier studies as a major 5S rRNA transcript. It differs from the reference sequence used in this study by a single T→C substitution at position 96, resulting in an A-C mispairing within the helix V of the 5S rRNA's secondary structure (Cloix *et al.*, 2002). The remaining variants display from one to 19 substitutions and/or single or multiple nucleotide insertions and deletions when compared to the reference sequence. The analysis of the effect of the sequence variations on the 5S rRNA secondary structure revealed that in the case of 90 variants, the nucleotide substitutions occur in the single-stranded regions. In the case of substitutions of nucleotides involved in base pairing, they are compensated to produce either another Watson-Crick or a G-U base pair. However, the folding of the majority of variants would produce secondary structures with up to 10 mispaired positions relative to the structure of the reference sequence.

5S rRNA derived short RNAs

To address the question of the expression of 5S rRNA variants in *Arabidopsis thaliana,* we qualitatively analyzed a large set of 942.4 millions of pooled short (18–30 nt long) RNAs from various tissues and developmental stages obtained from the next generation sequencing experiments. The sequencing reads were mapped to the target reference sequence of *Arabidopsis* 5S rRNA using *blast* allowing up to 30% of mismatches. To increase the stringency of the analysis, in the subsequent filtering steps only the reads that matched the target sequence along the entire length and were present in at least 500 copies in the whole set were retained.

Mapping and filtering resulted in a set of 460 unique sequences corresponding to 3.385 millions of individual reads. Among these sequences, 367 (3.175 million reads) exactly matched the reference sequence and accounted for 94% of all reads. The remaining 93 sequences (0.21 million reads) aligned to the reference 5S rRNA with one to four mismatches. The summary of the substitutions corresponding to these sequences and their positions on the 5S rRNA secondary structure are shown in Table 1 and Fig. 1A. With a few exceptions, the majority of the sequence variations represented by these reads affect positions involved in base pairing and their presence in the sequence could interfere with the formation of the correct secondary structure of 5S rRNA.

Individual substitutions represented by all of these variants can be found in genomic sequences. However, a comparison of the 5S rDNAs from Rfam with the altered positions in the short RNA sequencing reads revealed only nine unique gene sequences consistent with specific combinations of substitutions (Table 2). All of these genes represent variants with C at position 96, replacing U that is involved in base-pairing with A80. The
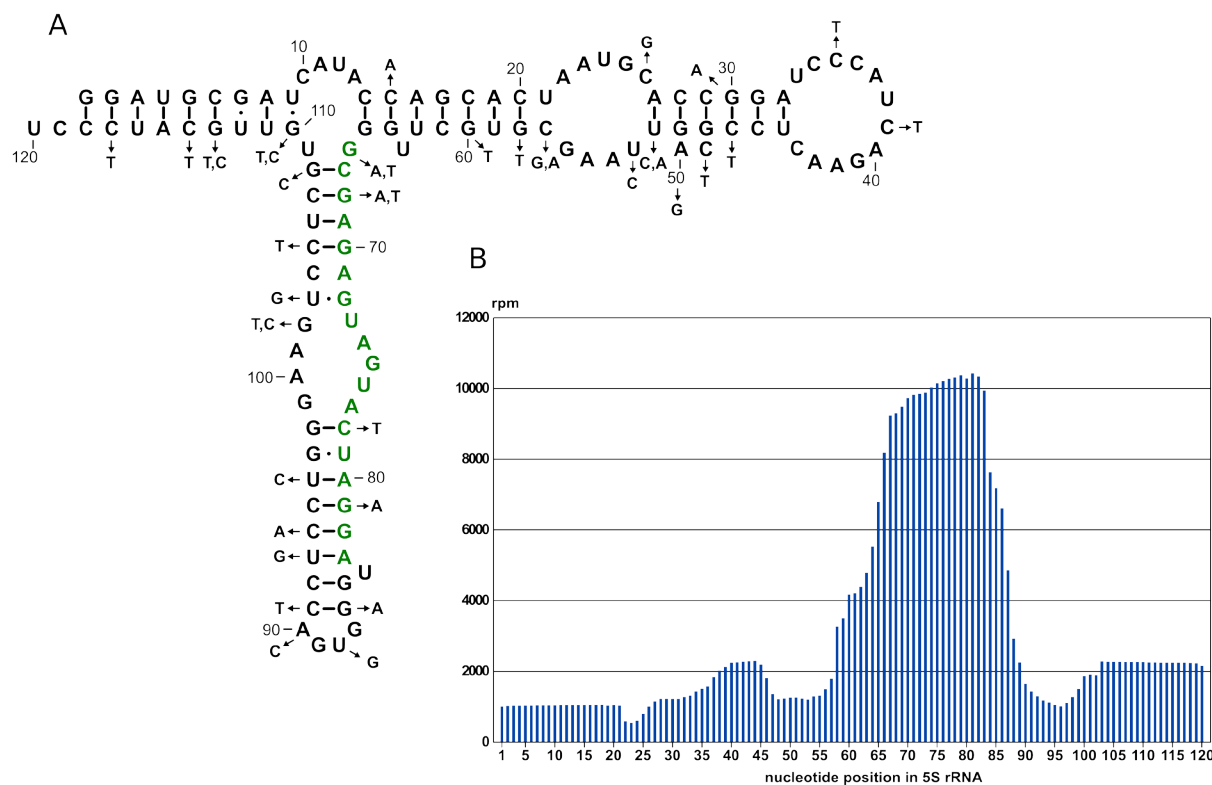
Table 1. Nucleotide substitutions in the 5S rRNA identified in the short RNA sequencing data.
The number of reads containing the substitution and a normalized value (reads per million) are shown in the second column

1

Substitution(s)	Number of reads (reads / 10^6)
35T	568 (2.55)
39T	666 (2.99)
47T	7845 (35.12)
49T	741 (3.33)
52C	1027 (4.61)
53C	972 (4.37)
57G	2428 (10.90)
60T	4547 (19.42)
66A	1081 (9.35)
66T	587 (2.64)
68A	540 (2.43)
68T	2435 (10.53)
78T	1731 (6.77)
96C	164610 (733.73)
102T	2486 (11.16)
102C	1542 (6.92)
103G	1400 (6.25)
105T	603 (2.71)
108C	701 (3.15)
110T	631 (2.83)
110C	514 (2.31)
113T	540 (2.43)
113C	576 (2.59)
114T	1378 (6.19)
117T	2886 (11.68)
78T, 96C	1765 (7.93)
15A, 26G, 30A	729 (3.27)
48A, 51A, 57A, 58T	1529 (6.86)
48A, 49G, 50G, 51A	1447 (6.50)
81A, 86A, 90C, 91T	2002 (8.99)
88G, 92G, 93G, 94A	553 (2.48)

78T, 47T and 60T substitutions result in wobble G-U base pairs replacing the Watson-Crick G-C pairs in the reference structure. The remaining positions are located within loops and their changes should not alter overall secondary structure, although they could affect the ability of 5S rRNA to interact with proteins or other RNAs.

The heterogeneity of *Arabidopsis thaliana* 5S rRNAs was previously analyzed using sequencing with the use of a reverse transcriptase (Cloix *et al.*, 2002). The results of this research demonstrated that there are several sequence variants of 5S rRNA expressed, including one major (the most abundant) and several minor 5S rRNA species. All of these sequences had shown a high degree of similarity and the differences in the nucleotide sequences between the major and each of the minor 5S rRNAs were limited to single or double nucleotide substitutions. The heterogenous positions identified by this

Figure 1 (A) The nucleotide sequence of the reference 5S rRNA (Barciszewska *et al.*, 2004) with substitutions identified in the short RNA sequencing data. (B) A profile of the coverage of the 5S rRNA sequence by sequencing reads. The coverage values are given in reads per million relative to the whole set of sequences of the length of 18–30 nt. The positions corresponding to the peak in the coverage plot are shown in green on the structure diagram.

Table 2. The 5S rDNA sequences from the Rfam database compatible with substitutions identified within the pool of short RNAs.
A representative GenBank accession number and sequence coordinates of the 5S RNA coding sequence are given. The third column shows a number of individual sequences corresponding to the representative sequence in the Rfam database.

Representative accession number: coordinates	Compatible substitution combination	Sequences in Rfam
CP002688.1:11186302-11186420	96C	161
AY130740.1:319-437	78T, 96C	5
AF335815.1:136-254	39T, 47T, 96C	3
CP002687.1:3058158-3058276	39T, 96C	4
AY130747.1:190-308	35T, 53C, 96C	1
CP002688.1:11184796-11184914	53C, 96C	1
AF331003.1:1-119	52C, 96C	2
AY130645.1:319-437	47T, 96C	6
M65137.1:220-338	60T, 96C	2

approach included 21C, 40T, 47T, 51A, 52C, 53C, 56T, 59C, 64A, 78T, 91T, 93A, 93C, 94T, 96T and 101T. It has been demonstrated that the contribution of the minor 5S rRNA species to the total pool of 5S rRNAs differs between tissues and developmental stages. In seeds and developing seedlings, the minor transcripts account for ~13% of 5S rRNA, whereas in adult tissues they constitute only 3% of the total 5S rRNA pool (Cloix *et al.*, 2002).

Transcriptionally active 5S rDNA units encoding the major and minor 5S rRNA transcripts were found only in the clusters located on chromosome 4, and in the large locus of chromosome 5 (Cloix *et al.*, 2002). It has been estimated that the repeats producing the major 5S rRNA variant account for about 10% of all 5S rDNA repeats, and only for 20% of the genes present in the active clusters on chromosomes 4 and 5 (Cloix *et al.* 2002). An analysis of the 5S changes in the rDNA chromatin structure and epigenetic modifications suggested that the transcriptional control of 5S rRNA genes is associated with DNA methylation and specific modifications of histone H3 (H3K9 acetylation and H3K4 methylation) changes during development (Mathieu *et al.*, 2003). Altered methylation patterns were shown to be responsible

for the transcriptional regulation of the 5S rDNA units encoding minor variants of the 5S rRNA (Vaillant *et al.*, 2008).

Interestingly, the major 5S rRNA variant described by Cloix and coworkers (2002), differing only by the 96C substitution when compared with the reference sequence, does not produce the majority of the short RNAs identified in our analysis. The sequencing reads supporting this sequence variant account for ~25% of all reads mapping to position 96, whereas most of the reads contain C at this position and are consistent with the reference sequence. In our sequencing data, we also found reads supporting six substitutions consistent with the minor 5S rRNA species, including 47T, 52C, 53C, 56T, 78T and 96T. However, we were unable to identify the remaining 10 substitutions. Changes at these positions were observed in a fraction of reads, but they were always associated with other substitutions that were not reported in the minor variants. It has to be noted however, that our analysis is based on short RNAs and, as such, will promote identification of stable fragments derived from the transcribed 5S rRNA molecules. Therefore, the absolute representativeness of each of the 5S rRNA variants may be biased. Such a bias may account for the apparent discrepancy between our results and the results of the sequencing of the RT-PCR products reported previously (Cloix *et al.*, 2002) and concerns primarily the relative abundance of the major variants differing at position 96.

Another interesting feature of the data from the short RNA sequencing is the profile of a coverage of particular positions within 5S rRNA with sequencing reads (Fig. 1B). One can notice that the coverage is not uniform and that the reads mapping to the positions 67–83 which correspond to the 5' portions of helix IV, loop E and helix V in the 5S rRNA secondary structure, are significantly more stable than fragments from the 5'- and 3'-flanking regions. The functional significance of this fragment, if any, poses an interesting question for future investigations.

The question if the heterogeneity of 5S rRNAs is somehow associated with their functions remains unresolved. A new regulatory role of the 5S rRNA complexes with ribosomal proteins was identified in mammalian cells, where the 5S RNP complex consisting of 5S rRNA and L5 and L11 ribosomal proteins regulates the activity of p53, providing a link between cell proliferation and biogenesis of the ribosomes (Sloan *et al.* 2013). Also, little is known about a possible functional impact of incorporation of different 5S rRNA variants into the ribosomes.

Acknowledgments

REFERENCES

Barciszewska MZ, Erdmann VA, Barciszewski J (1994) A new type of RNA editing. 5S ribosomal DNA transcripts are edited to mature 5S rRNA. *Biochem Mol Biol Int* **34**: 437–448

Baum BR, Edwards T, Johnson DA (2013) What does the 5S rRNA multigene family tell us about the origin of the annual Triticeae (Poaceae)? *Genome* **56**: 245–266. doi: 10.1139/gen-2012-0195

Baum BR, Johnson DA Baum BR, Johnson DA (1994) The molecular diversity of the 5S rRNA gene in barley (*Hordeum vulgare*). *Genome* **37**: 992–998

Bellavia D, Dimarco E, Naselli F, Caradonna F (2013) DNA-methylation dependent regulation of embryo-specific 5S ribosomal DNA cluster transcription in adult tissues of sea urchin *Paracentrotus lividus*. *Genomics* **102**: 397–402. doi: 10.1016/j.ygeno.2013.08.001

Campell BR, Song Y, Posch TE, Cullis CA, Town CD (1992) Sequence and organization of 5S ribosomal RNA-encoding genes of *Arabidopsis thaliana*. *Gene* **112**: 225–228

Ciganda M, Williams N (2011) Eukaryotic 5S rRNA biogenesis. *Wiley Interdiscip Rev RNA* **2**: 523–533. doi: 10.1002/wrna.74

Cloix C, Tutois S, Yukawa Y, Mathieu O, Cuvillier C, Espagnol MC, Picard G, Tourmente S (2002) Analysis of the 5S RNA pool in Arabidopsis thaliana: RNAs are heterogeneous and only two of the genomic 5S loci produce mature 5S RNA. *Genome Res* **12**: 132–144. doi: 10.1101/gr.181301

Dimarco E, Cascone E, Bellavia D, Caradonna F (2012) Functional variants of 5S rRNA in the ribosomes of common sea urchin *Paracentrotus lividus*. *Gene* **508**: 21–25

Ford PJ, Southern EM (1973) Different sequences for 5S RNA in kidney cells and ovaries of *Xenopus laevis*. *Nat New Biol* **241**: 7–12

Fransz P, Armstrong S, Alonso-Blanco C, Fisher T C, Torres-Ruiz R A, Jones G (1998) Cytogenetics for the model system *Arabidopsis thaliana*. *Plant J* **13**: 867–876

Hariharan N, Reddy PS, Padayatty JD (1987) 5S-rRNA genes in rice embryos. *Plant Mol Biol* **9**: 443–451

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948

Mathieu O, Jasencakova Z, Vaillant I, Gendrel AV, Colot V, Schubert I, Tourmente S (2003) Changes in 5S rDNA chromatin organization and transcription during heterochromatin establishment in Arabidopsis. *Plant Cell* **15**: 2929–2939

Murata M, Heslop-Harrison JS, Motoyoshi F (1997) Physical mapping of the 5S ribosomal RNA genes in *Arabidopsis thaliana* by multicolor fluorescence in situ hybridization with cosmid clones. *Plant J* **12**: 31–37

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43**: D130–D137. doi: 10.1093/nar/gku1063

Peterson RC, Doering JL, Brown DD (1980) Characterization of two xenopus somatic 5S DNAs and one minor oocyte-specific 5S DNA. *Cell* **20**: 131–141

Sloan KE, Bohnsack MT, Watkins NJ (2013) The 5S RNP couples p53 homeostasis to ribosome biogenesis and nucleolar stress. *Cell Rep* **5**: 237–247. doi: 10.1016/j.celrep.2013.08.049

Szymanski M, Barciszewska M, Erdmann VA, Barciszewski J (1995) Are 5S ribosomal RNA primary transcripts edited? *Biochem Mol Biol Int* **37**: 413–422

Szymański M, Barciszewska MZ, Erdmann VA, Barciszewski J (2003) 5 S rRNA: structure and interactions. *Biochem J* **371**: 641–651

Szymanski M, Zielezinski A, Barciszewski J, Erdmann VA, Karlowski WM (2016) 5SRNAdb: an information resource for 5S ribosomal RNAs. *Nucleic Acids Res* **44**: D180–D183. doi: 10.1093/nar/gkv1081

Vaillant I, Tutois S, Jasencakova Z, Douet J, Schubert I, Tourmente S (2008) Hypomethylation and hypermethylation of the tandem repetitive 5S rRNA genes in Arabidopsis. *Plant J* **54**: 299–309. doi: 10.1111/j.1365-313X.2008.03413.x